Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Comparison of Sensibilities of Japanese and Koreans in Recognizing Emotions from Speech by using Bayesian Networks

Jangsik Cho, Shohei Kato, and Hidenori Itoh
Dept. of Computer Science and Engineering
Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466–8555, Japan
Email: {cho, shohey, itoh}@juno.ics.nitech.ac.jp

***Abstract*—The paper describes a comparison of the sensibility of recognizing emotions from human voices speaking Japanese and Korean. Our study focuses on the emotional elements included in the human voice, and our method uses Bayesian networks of prosodic features as models of Japanese's and Korean's sensibilities in recognizing emotions. The training datasets are prosodic features extracted from emotionally expressive voice data in the two languages. Our method makes the Bayesian network learn the dependence and its strength between nonverbal voice features and its emotion. We compare the sensibilities of emotion recognition from Japanese and Koreans speech by examining the cross-inference through two Bayesian networks with speech in the other language.**

***Index Terms*—comparison of sensibilities of Japanese and Koreans, Bayesian networks, emotion recognition from human voice**

I. Introduction

Human service robots and anthropomorphic software agents have been developed thanks to advances in information technology and robotics. In particular, man-machine interface and human-computer interaction (HCI) technologies, for expressive communication with humans, have been developed (e.g., [2], [3], [4], [19]). Communication involves not only conveying messages or instructions but also psychological interactions, such as comprehending mutual sentiments, sympathizing with the other person, and enjoying the conversation itself. Humans communicate expressively by recognizing their dialogist's emotions, having emotions, and expressing emotions. To communicate expressively, HCI technology requires several mechanisms to make up for its lack of human intelligence. Bayesian networks is a computer technology for dealing with the probabilities encountered in artificial intelligence. It is a probabilistic reasoning technique for circumstances involving uncertainty, and it is becoming increasingly important in research and applications of artificial intelligence (e.g., [1], [7], [15], [20]). In this study, we focused on a method for recognizing emotions. We have previously presented a Bayesian network-based method for detecting emotions in human voices [5]. The method focuses on prosodic features in emotionally expressive human voices and models the causal relationship between emotions and the features by using a Bayesian network. The Bayesian network built by our method can detect emotions from human voices expressing non-verbal information.

The world's communities are growing more and more interdependent. This trend has promoted economic cooperation, foreign trade, immigration, etc. In light of this trend, there has been a large increase in opportunities for cross-cultural exchange, and thus, mutual understanding of emotions in order to comprehend or sympathize with speakers of different languages is becoming even more important.

Besides vocal expression, facial expression is an important element of emotional expression. According to Paul Ekman's study [10] of human faces, people in many different cultures innately share facial expressions conveying six basic emotions (anger, sadness, disgust, fear, surprise, and happiness). With respect to emotional expression of the human voice, however, there has been almost no research comparing vocal emotional expressions of people who speak different languages and have different cultural backgrounds. Although there are several reports on detecting emotions in human voice (e.g., [14], [8], [18]), they deal with a specific language. We chose Japan and Korea as the two different cultures for our study. The grammars of Japanese and Korean show some similarity: they have the same word order, and nominatives are indicated by particles. On the other hand, speakers of these languages can't understand the other language because they have different phonologies, vocabularies, and writings.

In this paper, we model the sensibilities that Japanese and Korean have for emotional voices by learning Bayesian networks that can detect emotions in emotional voices of native Japanese and Koreans. We then compare the sensibilities of emotion recognition from speech between Japanese and Korean by examining the cross-inference through two Bayesian networks with speech in the respective foreign language.

II. Bayesian Network

A Bayesian network (BN) is a graphical structure that allows us to represent and reason about uncertain domain [15]. The graph structure is constrained to be a directed acyclic graph (or simply dag). A node in a Bayesian network represents a set of random variables from the domain. A set of

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

TABLE I
CONDITIONAL PROBABILITY TABLE FOR X_i

$p(X_i) = y_1 \| Pa(X_i) = x_1$	...	$p(X_i) = y_1 \| Pa(X_i) = x_m$
....	...	....
$p(X_i) = y_n \| Pa(X_i) = x_1$	...	$p(X_i) = y_n \| Pa(X_i) = x_m$

directed arcs (or links) connects pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node.

Most commonly, BNs are representations of joint probability distributions. Consider a BN containing n nodes, X_1 to X_n, taken in that order. A particular value in the joint distribution $P(X_1, ..., X_n)$ is calculated as follows:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(x_i|Pa(X_i)), \tag{1}$$

where $Pa(X_i) \subseteq \{X_1, ..., X_{n-1}\}$ is a set of parent nodes of X_i. This equation means that node X_i is dependent on only $Pa(X_i)$ and is conditionally independent of nodes except all nodes preceding X_i.

Once the topology of the BN is specified, the next step is to quantify the relationships between connected nodes. Assuming discrete variables, this is done by specifying a conditional probability table (CPT). Consider that node X_i has n possible values $y_1, ..., y_n$ and its parent nodes $Pa(X_i)$ have m possible combinations of values $x_1, ..., x_m$. The conditional probability table for X_i is as shown in Table I.

Once the topology of the BN and the CPT are given, we can do the probabilistic inference in the BN by computing the posterior probability for a set of query nodes, given values for some evidence nodes. Belief propagation (BP) proposed in [20] is a well-known inference algorithm for singly connected BNs, which have a simple network structure called a polytree. Assume that X is a query node and there is a set of evidence nodes E (not including X). The task of BP is to update the posterior probability of X by computing $P(X|E)$.

In the most general case, the BN structure is a multiply connected network, where at least two nodes are connected by more than one path in the underlying undirected graph. In such networks, the BP algorithm does not work; instead several enhanced algorithms, such as junction tree [13] logic sampling [12] and loopy BP [17], are used as exact or approximate inference methods.

In this study, we intend to reduce the scale of the knowledge base and computational cost dramatically by utilizing Bayesian networks for representing knowledge used in emotion detection.

III. EMOTION INFERENCE ENGINE

We focus on the prosodic features of the speaker's voice as a cue to what emotion he or she expresses. This section describes a BN modeling for this problem.

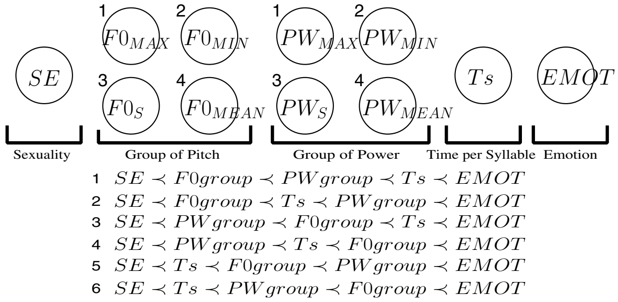

SE
1
$F0_{MAX}$
2
$F0_{MIN}$
3
$F0_S$
4
$F0_{MEAN}$
1
PW_{MAX}
2
PW_{MIN}
3
PW_S
4
PW_{MEAN}
Ts
$EMOT$
Sexuality
Group of Pitch
Group of Power
Time per Syllable
Emotion
1 $SE \prec F0group \prec PWgroup \prec Ts \prec EMOT$
2 $SE \prec F0group \prec Ts \prec PWgroup \prec EMOT$
3 $SE \prec PWgroup \prec F0group \prec Ts \prec EMOT$
4 $SE \prec PWgroup \prec Ts \prec F0group \prec EMOT$
5 $SE \prec Ts \prec F0group \prec PWgroup \prec EMOT$
6 $SE \prec Ts \prec PWgroup \prec F0group \prec EMOT$

Fig. 1. Possible variable orders of node groups

A. Voice Data

Voice data for learning the BN should be expressive of emotions. We used segments of Japanese and Korean voice samples that were spoken emotionally by actors and actresses in films, TV dramas, and so on. We labeled all segments with five emotional labels (anger, sadness, disgust, surprise or happiness). We extracted voice samples from many and unspecified actors and actresses, and Japanese and Korean collected sound data in their native language.

B. Features Extraction

Voice has three components: prosody, tone, and phoneme. It became obvious from reviewing past research that the prosodic component is the most relevant to emotional expressions [8] [21]. As attributes of voice data, we chose three prosodic attributes: energy, fundamental frequency and duration as the acoustic parameters for BN modeling. Prosodic analysis was done on 11 ms frames passed through a Hamming window extracted from voice waveforms sampled at 22.05 kHz.

The attributes of energy, maximum energy (PW_{MAX}), minimum energy (PW_{MIN}), mean energy (PW_{MEAN}) and its standard deviation (PW_S) are determined from the energy contours for the frames in a voice waveform. The attributes of fundamental frequency, maximum pitch ($F0_{MAX}$), minimum pitch ($F0_{MIN}$), mean pitch ($F0_{MEAN}$) and its standard deviation ($F0_S$) are determined from short time Fourier transforms for the frames in a voice waveform. As the attribute concerning duration, we measure the average duration per a single syllable(Ts). Then we added the attribute of the speaker's sexuality (SE).

The goal attribute ($EMOT$) and the above nine prosodic feature values and speaker's sexuality (total eleven attributes) were assigned to the nodes of the BN model.

C. Discretization of Feature

The section describes the discretization of extracted prosodic features. We considered Bayesian networks with discrete and multinominal variables only. In order to learn the discrete causal structure of the BN model, all prosodic features were converted into discrete values. The thresholds to discretize continuous values were determined from the distribution of the prosodic features extracted from the training voice samples.

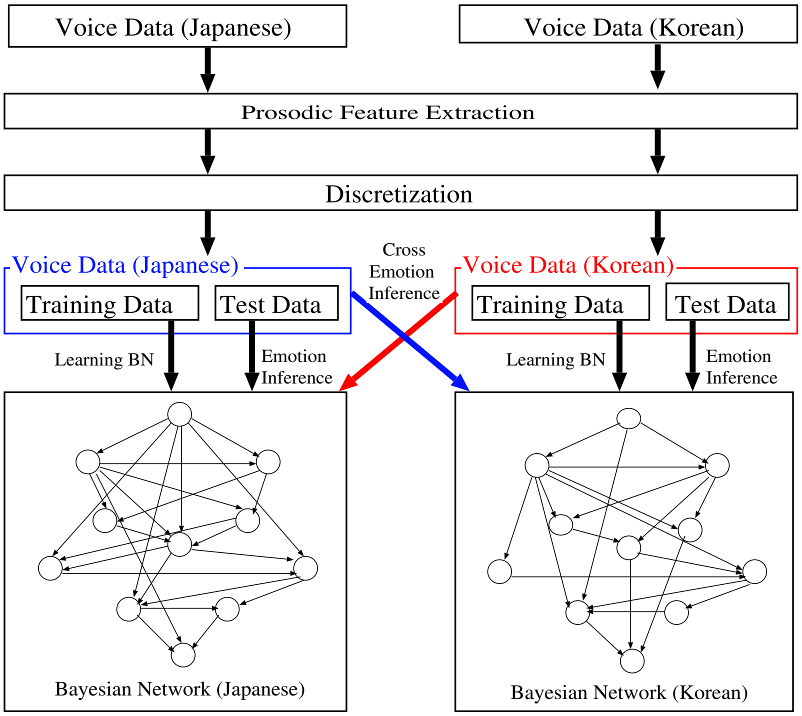


Fig. 2. Cross-inference for emotion detection from Japanese and Korean speech data

D. *Learning BN Structure*

The section describes how to specify the topology of the BN model for emotion detection and to parametrize CPT for connected nodes. The emotion detection BN modeling is to determine the qualitative and quantitative relationships between the output node containing the goal attribute (emotions) and nodes containing prosodic features. We chose a model selection method based on the Bayesian information criterion (BIC) [11], which has information theoretical validity and is able to learn a high prediction accuracy model through avoidance of over-fitting to training data.

Let M be a BN model, θ_M be a parameter representing M, and d be the number of parameters. M is evaluated by the BIC of M, defined as

$$BIC(\hat{\theta}_M, d) = -2\log P(D\,|\,\hat{\theta}_M) + d\log N, \tag{2}$$

where D is training samples, and $P(D\,|\,\theta_M)$ is the likelihood of D given θ_M; $\hat{\theta}_M$ is the parameter representing M giving the maximum likelihood (ML) estimate; and N is the number of the samples. If D is partially observed, expectation maximization (EM) algorithm [9] is utilized for estimating θ_M asymptotically with incomplete data in the training samples.

In this paper, as the knowledge for emotion detection, we made a BN model that maximizes BIC for training voice data, as a model of sensibility on emotion recognition from voice. We used K2 [6] [7] as the search algorithm. K2 needs a pre-selected variable order. We thus considered every possible permutation of three node groups: PW, $F0$ and Ts, such that node SE preceded all others shown in Fig. 1.

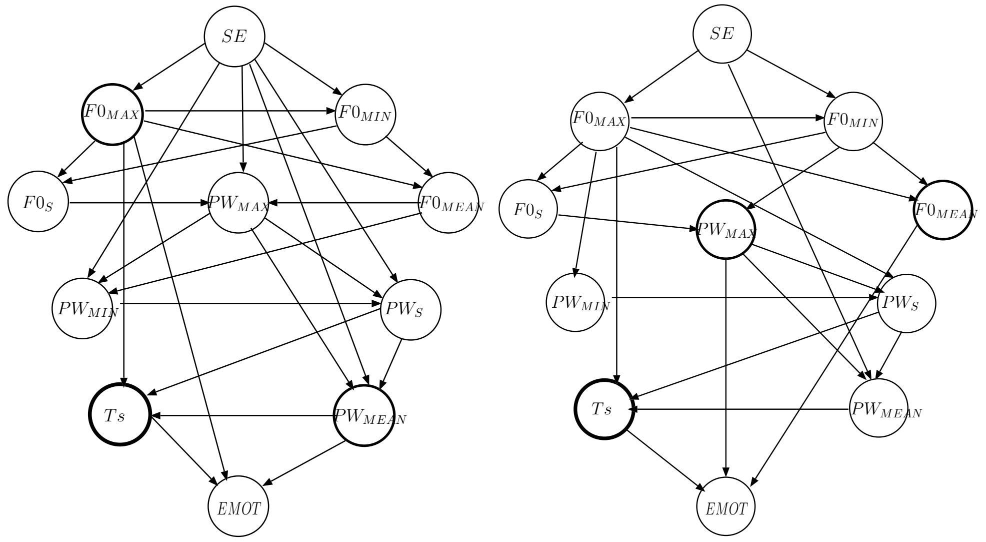
The BN has no verbal information: it only has prosodic features and speaker's sexuality. With respect to the native language, verbal information is often dominant in the emotion recognition from speech, but it is of no use for the foreign language. We used BNs composed of non-verbal information to enable a pure comparison of native and foreign languages.

E. *Inference Algorithm*

The topology of the BN is often multiply connected when there is a complicated relationship between variables. We chose junction tree [13] as the inference algorithm with BNs, it is an exact inference algorithm in multiply connected BNs. It is efficient clustering inference algorithms. Clustering inference algorithms transform the BN into a probabilistically equivalent polytree by merging nodes and removing multiple paths between two nodes along which evidence may travel.

IV. Comparison of Sensibilities through Cross-Language Emotion Inference

The section describes an experiment comparing the sensibilities of emotion recognition of Japanese and Koreans by using the Bayesian approach. Fig. 2 shows an overview of the experiment. First, we collected 500 segments of voice waveforms in Japanese and Korean (1000 segments in total) and labeled them with five emotions, as described in Section III-A. Then, we extracted nine prosodic features and the speaker's



(a) A BN_{JP} learned from Japanese voices

(b) A BN_{KR} learned from Korean voices

Fig. 3. Bayesian network structure learned from training data

sexuality in each of the segments and assigned them to the attributes, as described in Section III-B. The prosodic analysis used the Snack sound toolkit [22]. After that we randomly selected 400 samples from Japanese and Korean (800 in total) as training data and discretized their attributes into four values. We determined the threshold for discretization on the basis of the idea of even-sized chunks; that is, each discrete value covers 25% of the training data. We then modeled the BNs for Japanese and Korean by changing six variable orders (see Fig. 1) with Bayes Net Toolbox [16].

Fig. 3 shows the results of learning: BN for Japanese (BN_{JP}) and BN for Korean (BN_{KR}) with the variable order $SE \prec F0_{MAX} \prec F0_{MIN} \prec F0_S \prec F0_{MEAN} \prec PW_{MAX} \prec PW_{MIN} \prec PW_S \prec PW_{MEAN} \prec Ts \prec EMOT$, where both the BNs have the highest accuracy rates for their native voice samples. First, we compared the parent nodes of $EMOT$ between the BNs because these nodes strongly influence emotion inference. The parent nodes of $EMOT$ are $F0_{MAX}$, PW_{MEAN}, and Ts in BN_{JP} and are $F0_{MEAN}$, PW_{MAX}, and Ts in BN_{KR}. The results indicate roughly that the three prosodic features ($F0$, PW, Ts) are largely related with emotion inference in Japanese and Korean. Concerning the fundamental frequency ($F0$), Japanese sensibilities depend on the maximum value and Korean sensibilities depend on the average value. Concerning energy (PW), Japanese sensibilities depend on the average value and Korean sensibilities depend on the maximum value.

We then conducted two experiments on emotion inference: detecting emotions from native speech and from foreign speech. The first experiment attempted to confirm the effectiveness of the two BNs as sensibility models of Japanese and Korean. The second experiment was to enable a comparative discussion of sensibilities between Japanese and Koreans.

TABLE II
ACCURACY RATES OF EMOTION INFERENCE IN THE NATIVE LANGUAGE

		Accuracy Rates [%]			
		BN		PCA	
Language		Japanese	Korean	Japanese	Korean
Emotion	Anger	70	65	90	30
	Sadness	55	75	10	35
	Disgust	60	60	50	50
	Surprise	40	50	50	10
	Happiness	50	50	5	30
Total		55	60	41	31

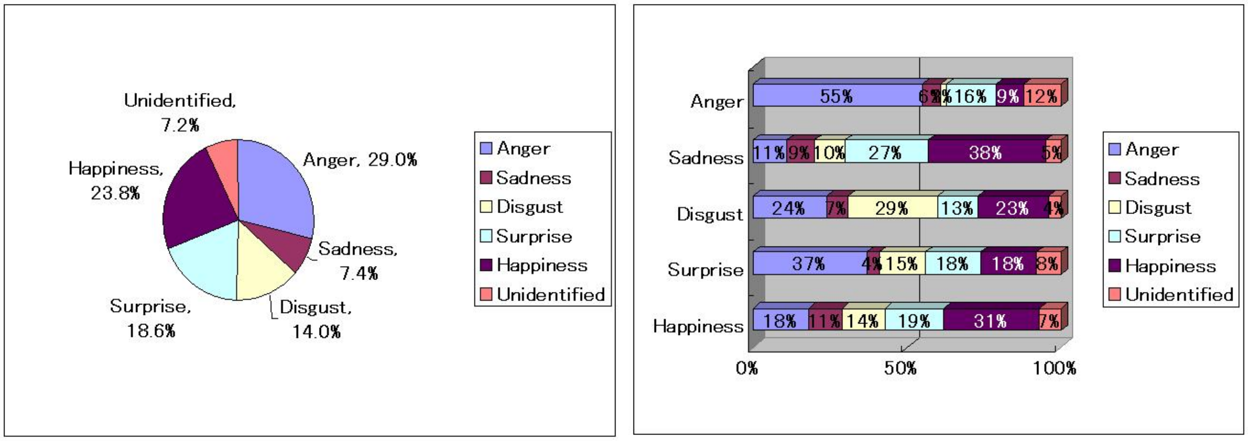
TABLE III
ACCURACY RATES OF EMOTION INFERENCE IN THE FOREIGN LANGUAGE

		Accuracy Rates [%]	
BN		BN_{JP}	BN_{KR}
Testdata		Korean	Japanese
Emotion	Anger	55	22
	Sadness	9	20
	Disgust	29	33
	Surprise	18	10
	Happiness	31	29
Total		28.4	22.8

A. Emotion Inference from Voices in the Native Language

We converted prosodic features of the remaining 100 samples in Japanese and Korean (200 in total) into discrete values by using the same thresholds for the training data and then examined the inference performance of each of the BN models shown in Fig 3. The left side of Table II shows the results. The BNs had accuracy rates of inference higher than 50%, except for Japanese surprise, and the total accuracy rates are higher than 55% in both Japanese and Korean.

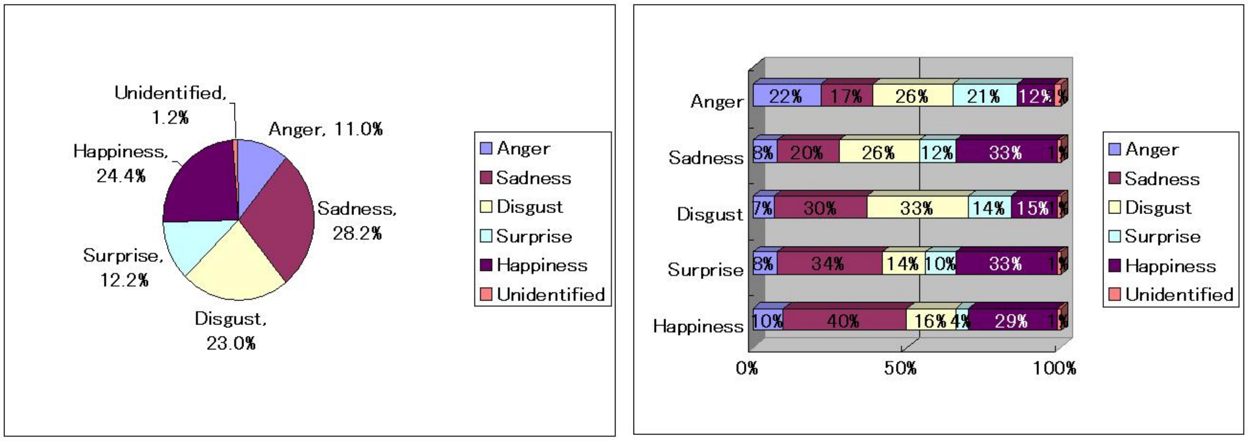
For comparison, we used principal component analysis (PCA) using ten features of training data and a classification based on the linear discriminant in a four-dimensional hyper-



(a) Inference rates

(b) Detailed inference rates of each emotion

Fig. 4. Detailed emotion inference rates for Korean voice using BN_{JP}



(a) Inference rates

(b) Detailed inference rates of each emotion

Fig. 5. Detailed emotion inference rates for Japanese voice using BN_{KR}

plane using four PCs because the accumulated contribution relevance was more than 70%. The right side of Table II shows the results. The BN for Korean had accuracy rates higher than the PCA for all emotions. The BN for Japanese had accuracy rates higher than the PCA, except for anger and surprise. The PCA for Japanese had very high accuracy rate for anger. To infer a specific emotion accurately is, however, totally insignificant, unless the BN can adequately infer all other emotions. The emotion inference has to get high average accuracy rates for all emotions. Note that the BNs for Japanese and Korean have total accuracy rates higher than the PCAs.

From now on, our discussion will proceed under the assumption that the BNs for Japanese and Korean reflect their sensibilities of emotion recognition from voices. With respect to each emotion, both BNs had high accuracy rates for anger. The BN for Korean had high accuracy rate for sadness as well. Both BNs had lower accuracy rates for surprise and happiness. These results suggest that Japanese and Koreans easily recognize anger and Koreans easily recognize sadness in their own native speech, and that it is slightly difficult for Japanese and Korean to recognize surprise and happiness in their own native speech.

B. Emotion Inference from Voices in a Foreign Language

To compare the sensibilities of Japanese and Koreans in recognizing emotions from voices, we examined the cross-emotion inference through two BNs with speech in the respective foreign language. The cross-emotion inference was done by giving 500 Korean voice samples to BN_{JP} and giving 500 Japanese voice samples to BN_{KR}. Table III shows the resulting accuracy rates. The accuracy rates for the respective foreign languages (see Table III) are lower than those for the native languages (see Table II). BN_{JP}'s accuracy rate for Korean anger was higher than 50%. This result suggests that Japanese sensibilities can easily recognize anger from Korean speech. BN_{KR}'s accuracy rate for Japanese disgust was higher than those of other emotions. The result suggests that Korean sensibilities can fairly easily recognize disgust from Japanese speech. With respect to sadness and surprise, both BN_{JP} and BN_{KR} had accuracy rates not more than 20%. The results suggest that it is difficult for Japanese and Koreans to recognize sadness and surprise from the other's speech.

Fig. 4a shows the inference rates of emotions using BN_{JP} with Korean voice samples. For example, 145 Korean voice samples (29%) are recognized as anger. The figure indicates that BN_{JP} recognizes most Korean voices as expressing anger or happiness and a few Korean voices as expressing sadness or disgust. Fig. 4b shows the detailed inference rates of each emotion. For example, 37 surprised voice samples (37%) are recognized as anger. The figure indicates that most Korean angry voices are recognized correctly but that many Korean surprised voices are mis-recognized as angry. The figure also indicates that many Korean sad voices are mis-recognized as happy. These results suggest that Japanese sensibilities often recognize Korean voices as angry and they often mis-recognize Korean surprise and sadness as anger and happiness, respectively.

Fig. 5a shows the inference rates for BN_{KR} on Japanese voice samples. For example, 141 samples (28.2%) from Japanese voice samples are recognized as sad. The figure indicates that most Japanese voices are recognized as sad or happy and a few Japanese voices are recognized as angry or surprised. Fig. 5b shows the detailed inference rates of each emotion. For example, 40 happy voice samples (40%) are recognized as sad. The figure indicates that lots of Japanese happy voices and surprised voices are mis-recognized as sad and lots of Japanese sad and surprised voices are mis-recognized as happy. These results suggest that Korean sensibilities often recognize Japanese voices as expressing sadness or happiness and they often mis-recognize Japanese happiness and surprise as sadness. These results also suggest that Korean sensibilities often mis-recognize Japanese sadness and happiness.

V. Conclusion

This paper described a comparative study using a Bayesian approach of sensibility of recognizing emotions of speech in different languages. This paper proposed a Bayesian-based method of detecting emotions from speech. The method uses a Bayesian network model of prosodic features of voices. We chose Japan and Korea as two different cultures, and modeled the sensibilities that Japanese and Koreans have for emotional voices by using Bayesian networks. We then compared the sensibilities of Japanese and Koreans, by examining the cross-inference using two Bayesian networks with speech in the respective foreign language. The experimental results showed that Japanese and Korean use different emotion expressions in speech. Furthermore, we found that Japanese recognize a lot of Korean voices as expressing anger and Korean recognize a lot of Japanese voices as expressing sadness. This result partially corresponds to the national sensibilities of Japanese and Koreans.

To improve the emotion inference ability of BNs for native languages, we will add attributes and solve the problem of individual differences between voice samples. We will also compare the cross-emotion inferences between BNs and the results of questionnaires to be given to Korean and Japanese people. We think that this study's results will be applicable to the development of an intelligent tool that helps to smooth communication between persons of different cultures.

References

[1] T. Akiba and H. Tanaka, "A Bayesian approach for user modelling in dialog systems", In 15th International Conference of Computational Linguistics, pp.1212-1218, 1994.

[2] C. Breazeal, "Designing socialable robots", MIT Press, Cambridge, 2002.

[3] R. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. Williamson, "The Cog project: building a humanoid robot", In C. L. Nehaniv, editor, Computation for Metaphors, Analogy and Agents, Lecture Notes in Artificial Intelligence, Vol. 1562, Springer-Verlag, 1999.

[4] W. Burgard, A. b. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner and S. Thrun, "The interactive museum tour-guide robot", Proc. of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), 1998

[5] J. S. Cho, S. Kato, and H. Itoh, "Bayesian-Based Inference of Dialogist's Emotion for Sensitivity Robots", 16th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN2007), pp.792-797, 2007.

[6] G. F. Cooper and E. Herskovits, "A Bayesian method for constructing Bayesian belief networks from databases", pp.86-94, 1991.

[7] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data, Machine Learning", 9,pp.309-347, 1992.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine", 18(1), pp.32-80, 2001.

[9] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royals Statistical Society", B 39,pp.1-38, 1977.

[10] P. Ekman and W. V. Friesen, "Unmasking the Face", Prentice-Hall 1975.

[11] G. Schwarz, "Estimating the dimension of a model", Annals of Statistics 6(2), pp.461-464, 1978.

[12] M. Henrion, "Propagating uncertainty in Bayesian networks by logic sampling, Uncertainty in Artificial Intelligence", 2, pp.149-163, 1988.

[13] F. V. Jensen, "Bayesian Networks and Decision Graphs", Springer-Verlag, 2001.

[14] Y. Kitahara and Y. Tohkura, "Prosodic Control to Express Emotions for Man-Machine Speech Interaction", IEICE Trans. Fundamentals, Vol.E75-A, No.2, pp.151-163, 1992

[15] K. B. Korb and A. E. Nicholson, "Bayesian Artificial Intelligence", Chapman & Hall/CRC, 2003.

[16] K. P. Murphy, Bayes Net Toolbox, http://www.cs.ubc.ca/˜murphyk/Software/BNT/ bnt.html.

[17] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference, an empirical study", pp.467-475, 1999.

[18] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models; Speech Communication", Vol.41, No.4, pp.603-623, 2003

[19] V. I. Pavlović, R. Sharma, T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: A review", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, pp. 677-695, 1997.

[20] J. Pearl, "Probabilistic Reasoning in Intelligent Systems", Morgan Kaufmann, 1988.

[21] K. R. Scherer, T. Johnstone, and G. Klasmeyer, Vocal expression of emotion, R. J. Davidson, H. Goldsmith, K. R. Scherer eds., Handbook of the Affective Sciences, Oxford University Press, 433-456, 2003.

[22] K. Sjölander, The Snack Sound Toolkit, http://www.speech.kth.se/snack.